Lexical Networks in !Xung and Ju

Thesis

Presented in Partial Fulfillment of the Requirements for the BS Honors Research Distinction in the Undergraduate School of The Ohio State University

By

Syed-Amad Ashraf Hussain

Undergraduate Program in Computer Science Engineering

The Ohio State University

2018

Thesis Committee

Dr. Micha Elsner, Adviser

Dr. Eric Fosler-Lussier

Abstract

We investigate the lexical network properties of the large phoneme inventory Southern African languages Mangetti Dune !Xung and Ju|'hoansi as they compare to European languages. We conduct analyses over a range of lexicon sizes in search of disparities in the mental lexicons of the native speakers. We find no substantial disparity within these analyses so we continue on to simulate data ("pseudolexicons") with varying levels of phonotactic structure which find that the lexical network properties of !Xung and Ju diverge from European languages when fewer phonotactic constraints are retained. We conclude that lexical network properties are representative of an underlying cognitive structure which is necessary for efficient word retrieval and that the phonotactics of these languages are shaped by a selective pressure which preserve network properties within this cognitively useful range.

Acknowledgments

Thanks to Amanda Miller, Rory Turnbull, Philippa Shoemark and the OSU Phonies group. Funded by NSF 1422987.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

We investigate the lexical network properties (LNPs) of the Southern African languages Mangetti Dune !Xung and Ju|'hoansi (hereafter !Xung and Ju respectively) as they compare to European languages. The consonant phoneme inventories of !Xung and Ju are 87 and 89 respectively which are substantially larger than most European languages (Miller, 2016; Miller-Ockhuizen, 2003; Dickens, 1994; Maddieson, 2013). Many of these sounds are clicks, typologically rare sounds found mostly in Southern Africa. In !Xung and Ju, close to 90% of content words begin with an initial click. While these properties place !Xung and Ju distinctly apart from European languages at the phonemic level, we analyze their lexical networks (LNs) to determine whether their mental lexicons reflect these disparities.

In a LN, as shown in Figure 1, nodes represent words and edges between nodes represent minimal pairs (Vitevitch, 2008). Vitevitch (2008) argues that the high connectivity and tendency toward clustering found in the English language lexicon are important aids to word learning and retrieval; later work finds similar properties in other European language lexicons (Shoemark et al., 2016). Because !Xung and Ju have very large phoneme inventories, they might in principle have very different network properties from previously studied languages. Any given word might have far more minimally different neighbors; alternately, the words might be spread out more thinly across a wider phonemic space. We investigate whether the network properties of !Xung and Ju differ

from European languages. If not, what phonological properties of the language lead to this despite the large phoneme inventory?

Our initial analysis does not show substantial distinctions between the network properties of !Xung and Ju and European languages. We next look at these properties over a range of lexicon sizes. Because large lexicons for our languages are not available, we conduct these analyses on simulated data ("pseudolexicons") sampled from trigram models, following Gruenenfelder and Pisoni (2009). We compare our results against Shoemark et al. (2016) to again find no substantial difference from European languages. We then construct pseudolexicons with varying degrees of phonological structure (Turnbull and Peperkamp, 2016) and compare them against each other within each language. We show that !Xung and Ju are more susceptible to the loss of phonotactic structure than English. To determine what constitutes the bulk of this phonotactic structure, we create additional pseudolexicons that focus on the properties of !Xung and Ju. We find that pseudolexicons based on syllabic patterns and click locations move closer to the properties of the actual language but a disparity remains present. Overall, we find !Xung and Ju do not substantially differ in network properties when compared to European languages. However, when certain phonotactic properties are removed through the use of pseudolexicons, disparities between these languages and English arise, hinting at a greater reliance on phonotactics.

Figure 1: Example lexical network centered around the word “plan” (Turnbull and Peperkamp, 2016)

Previous Works

We conduct our analysis on LNs to derive cognitive and phonotactic conclusions. Vitevitch (2008) first presents this network model which assigns words as nodes and minimal pairs between these words as edges. He finds that lexical retrieval and language acquisition is aided by higher network density -- largely defined by the network properties of assortative mixing and average clustering coefficient.

Vitevitch (2008) and subsequent work on networks (Shoemark et al. 2016; Turnbull and Peperkamp, 2016; Stella and Brede, 2015), describe network structure in terms of four properties (Shoemark et al. 2016): Fraction in Largest Island is defined as the percent of the lexicon that is connected to the largest component, or island, in the network and characterizes the global connectivity of the network; Degree Assortativity Coefficient shows the tendency of nodes to be connected to other nodes with similar degrees where with higher values the central "hubs of the network are connected to one another (Newman and Girvan, 2003); Average Shortest Path Length (ASPL) averages the minimum number of hops it takes to get between any two nodes in the largest island, similar to the base concept of the game "Six Degrees to Kevin Bacon"; average Clustering Coefficient (CC) is defined as the number of edges that exist between neighbors divided by the number of possible edges between neighbors and can be thought of as "are my neighbors also neighbors with each other" or "do all my friends know each other?".

Later work on this model points out network statistics are affected by lexicon size, phoneme inventory size, word length distribution, and the inclusion of morphological

variants (Shoemark et al., 2016) Since these cannot all be controlled in cross-linguistic comparisons, an indirect comparison is necessary. The phonological properties of the language are used to generate pseudolexicons which are examined over several lexicon sizes. The trends for each language are then compared qualitatively against each other language (Shoemark et al., 2016).

Further work expands the use of pseudolexicons to determine the source of the network property statistics (Turnbull and Peperkamp, 2016). These pseudolexicons varied in how many, and which, phonotactic properties of the original language they retain. Turnbull and Peperkamp (2016) conclude that the typical range of values of average CC are intrinsic to all LNs, typical values of largest island size and ASPL are determined by phonological rules, and degree assortativity may reflect some higher-level organization principle within the lexicon. We employ a series of pseudolexicons which preserve various aspects of !Xung and Ju phonology to determine which phonological rules within these languages are responsible for preserving the typical values of largest island size and ASPL. We find that constraints on click placement and syllable structure can explain most, but not all the difference between randomly generated pseudolexicons and the real data.

Phonological Properties of !Xung and Ju

Ju|'hoansi and Mangetti Dune !Xung both belong to the Kx'a language family (formerly known as the Northern Khoisan branch of the Khoisan family). Ju|'hoansi is a member of the Southeastern branch of the Juu subgroup, while Mangetti Dune !Xung is a member of the Northern branch of the Juu subgroup, according to Sands' (2010) classification. The

complete sound inventory of Ju|'hoansi is provided in Dickens (1994) and Miller-Ockhuizen (2003). The complete sound inventory of Mangetti Dune !Xung is provided in Miller (2016). Ju|'hoansi has 89 consonants, 47 of which are click consonants, while Mangetti Dune !Xung contains 87 consonants, 45 of which are click consonants. Both languages also contain extremely large vowel inventories. There are only five contrastive vowel qualities, but there are many contrastive vocalic phonation types (modal, breathy, epiglottalized and glottalized), and the language also contrasts oral vs. nasal vowels. Nasality can combine with all the different phonation types, though there are some restrictions on which vowel qualities can combine with epiglottalization and nasalization. Both languages are tone languages. Each mora may bear one of 4 distinct tone levels with some restrictions on their co-occurrence (see Miller-Ockhuizen 2003), leading to 7 possible contrastive tone patterns that occur on content words. Over 90% of content words in both languages commence with a click consonant, while function words largely begin with a pulmonic (non-click) consonant.

Chapter 2. Preliminary Examination

In this preliminary examination, we implement LN models for our actual lexicons to establish baseline values for comparison against Shoemark et al. (2016).

Methodology

For this analysis, we create LNs using each of our 3 corpora directly. Our !Xung and Ju corpora contain 974 and 3733 words respectively and were obtained from field work (Biesele et. al. 2006; Miller et. al. 2008). Our English corpus is a list of the 974 highest frequency words from the Fisher corpus (David et. al., 2004). The English lexicon size was chosen to best compare against our !Xung lexicon.

We build each LN by assigning words as nodes and minimal pairs as edges. We build and analyze our networks using the python NetworkX package. From these networks, we derive the Fraction in Largest Island, Degree Assortativity, ASPL, and Average Clustering Coefficient. We then qualitatively compare these results to the values for the 8 reported European languages in Shoemark et al. (2016) to see if Ju and !Xung have substantially different properties. Due to their different lexicon sizes, Ju and !Xung are compared to different points on trend lines.

Results

The results and LN degree statistics are summarized within Table 1 and Table 2 of the appendix respectively. In general most of the network property statistics of !Xung and Ju do not differ greatly from the European language data presented in Shoemark et al. (2016) as each result has a relatively similar European language counterpart.

Chapter 3: Analysis 1

The preliminary examination did not show any definitive distinction when comparing !Xung and Ju against European languages but, as explained by Shoemark et al. (2016), we need to view LN data trends over several lexicon sizes. With the limited data available for these under-resourced languages, we cannot adjust the lexicon sizes. Instead we opt for the use of pseudolexicons derived from models based on the original lexicon as in Shoemark et al. (2016).

Methodology

To create pseudolexicons that most accurately capture the phonotactics of each language, we use a trigram model with Kneser-Ney smoothing (Turnbull and Peperkamp, 2016). Here, the probability of a given phoneme is conditioned on the probability of the preceding 2 phonemes. Kneser-Ney smoothing is applied to provide weight to distributions that might occur within the language but not be present within our lexicon (Turnbull and Peperkamp, 2016). Using these probabilities, we simulate "words" which follow the language's phonotactic properties but may not be actual words.

We train the trigram models using the KenLM Language Model Toolkit before using the SRI Language Modeling (SRILM) Toolkit to generate the simulated words (Heafield, 2011; Stolcke, 2002; Stolcke et. al. 2011). We generate pseudolexicons of size 2^{10}, 2^{11}, 2^{12}, and 2^{13} for each language (we did not generate a 2^{13} length pseudolexicon for English) and average relevant network statistics over 10 trials. Finally, we plot these network statistics and compare these trend lines to those generated by the European

Language pseudolexicon trend lines presented in Shoemark et. al. (2016).We follow the procedure from the preliminary examination to create and analyze each LN.

Results

We compare the network property trends over different pseudolexicon sizes to the results found in Shoemark et al. (2016). In an initial overview of Figure 2, we see that !Xung and Ju trend lines do not differ substantially from English. When looking at the corresponding charts in Shoemark et al. (2016), we see a steady rise in degree assortativity as lexicon size increases while the average CC generally remains the same. Figure 2 shows that our results follow these general patterns with average CC remaining steady and degree assortativity trending upwards. When comparing the fraction of nodes in the largest island, we notice the substantial upwards trend for !Xung and Ju mimic the results of Spanish or Portuguese in Shoemark et al. (2016). Finally, Figure 2 shows that the ASPL trends remain generally flat over different lexicon sizes. This is the largest disparity with Shoemark et al. (2016) in which most European languages trend upwards, however the pseudolexicon trend line for Portuguese is also generally flat.

The natural English trend line is provided to show the difference between pseudolexicons over a range versus their natural language counterpart. The divergence between natural English and pseudo-English can be explained by the fact that as the natural English lexicon increases in size, it begins to include rarer words which may be phonotactically unusual compared to more common English words. For example, these words may be longer than typical English words or contain rarer phonemes. Since larger pseudo-English lexicons are generated by extrapolating a smaller data set, these unusual

phonotactics are never introduced. Shoemark et. al. (2016) highlights that in certain languages, like Portuguese and Spanish, the rarer words are not phonotactically unusual, generating trend lines more similar to that of !Xung and Ju in Figure 2. However, since we only have limited lexicons of !Xung and Ju, we cannot be sure of the properties their larger lexicons might hold.

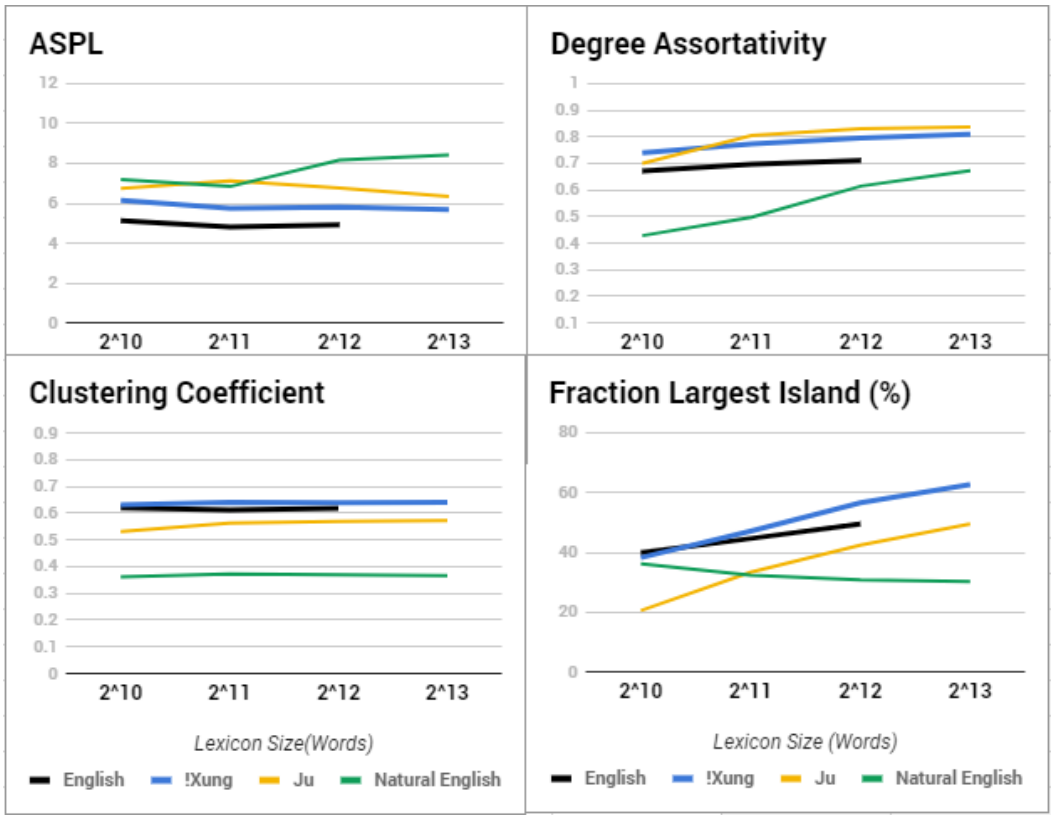

Figure 2: Trigram pseudolexicon network property values for English, !Xung, Ju as a function of lexicon size. Natural English data (not a pseudolexicon) is provided for comparison

Overall, most of the network property trends of !Xung and Ju do not differ greatly from the European language data presented in Shoemark et al. (2016).

Chapter 4: Analysis 2

Analysis 1, like the preliminary examination, did not show a definitive distinction between !Xung and Ju against European languages. This entails the question: what phonological properties allow !Xung and Ju to have similar LNPs to European languages despite their large phoneme inventory? In this analysis, we employ the methods used by Turnbull and Peperkamp (2016) to create pseudolexicons with varying levels of phonological structure. A comparison of these pseudolexicons highlights the phonotactic disparities between !Xung and Ju versus English.

Methodology

For each of our corpora, we generate the following pseudolexicons (Turnbull and Peperkamp, 2016): Uniform – randomly selects from the phoneme inventory; Zipfian – randomly selects from the phoneme inventory given a Zipfian distribution; Scrambled – scrambles the phonemes of a word in place; Bigram – like the previously mentioned trigram however it only accounts for the previous 1 phoneme; Trigram. We also create a Unigram pseudolexicon which randomly selects from the actual phoneme distribution. The pseudolexicons have the same word length distribution as the original lexicon. Examples of words from these pseudolexicons are shown in Table 3 within the appendix.

We compare the network properties of these pseudolexicons (averaged over 10 trials) within each language and examine any overarching trends cross-linguistically. We follow the procedure from the preliminary examination to create and analyze each LN.

Results

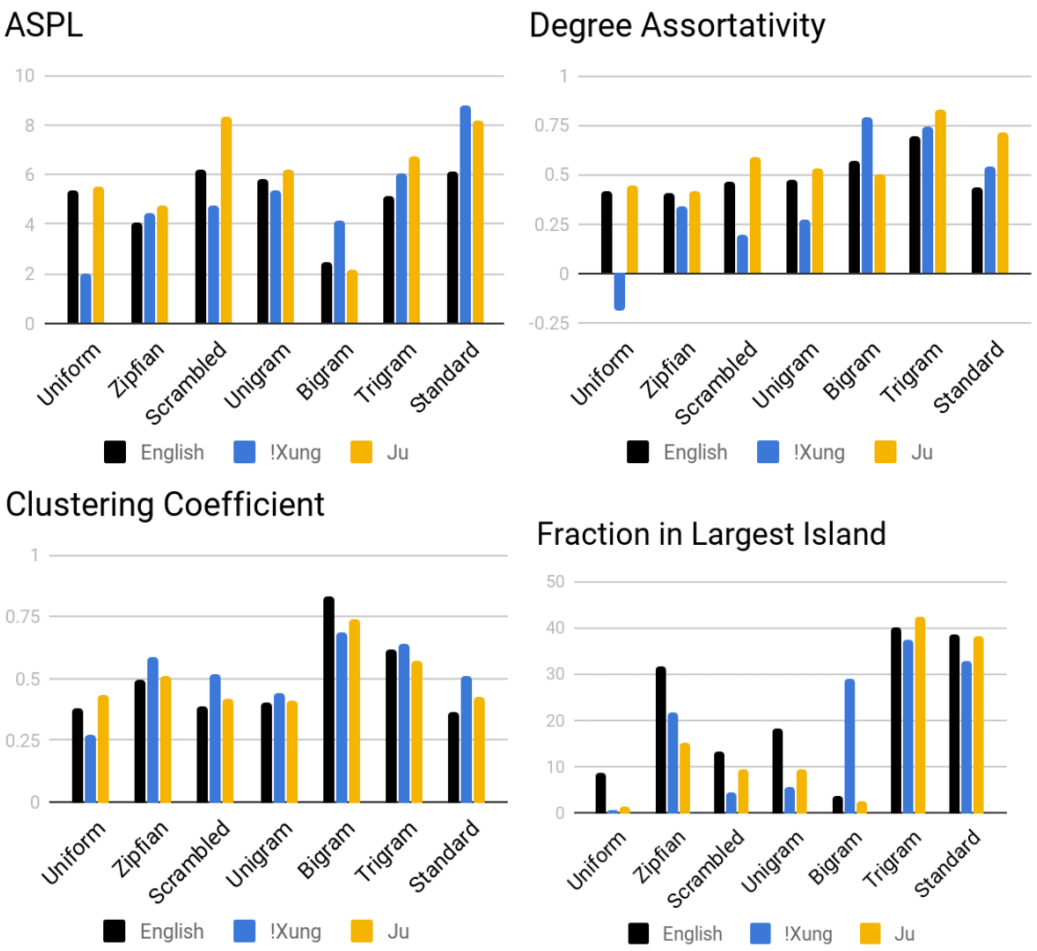

Figure 3: Network property values for English, !Xung and Ju over several different pseudolexicon models ordered based on how phonotactically similar they are to the natural language with right-most being the natural language itself.

We first compare the pseudolexicons of each language within themselves. Figure 3 (lower right) shows that only the trigram lexicon generates realistic sizes for the largest island. Other pseudolexicons are highly disconnected. This is especially the case for !Xung and Ju relative to English; for instance, the English uniform pseudolexicon (far left) has nearly 10% of the nodes in the largest island, while the other languages have essentially none. This disparity between languages is caused by the large phonemic

inventory which create fewer minimal pair matches when randomly sorted, as in the uniform, Zipfian, scrambled, and unigram pseudolexicons. The disparity begins to shrink as the pseudolexicons become more natural, suggesting disparities due to the large phonemic inventory are reduced by phonological structure and that phonotactic constraints on word forms in !Xung and Ju lead the lexicon to include more minimal pairs.

The other plots show the remaining network statistics; most of these are calculated on the nodes in the largest island, making them unreliable for most of the pseudolexicon types. For the lexicons with reasonable island sizes, the remaining three measurements are relatively close to the real values. Overall, the results suggest that only trigram pseudolexicons have enough phonotactic structure to ensure realistic network structures. In particular, bigrams have unrealistic measurements for average shortest path lengths and clustering coefficients for all languages. Again, however, the discrepancies between uniform lexicons and the real values are larger for !Xung and Ju than for English. These results emphasize the importance of phonotactic constraints for keeping the !Xung and Ju network properties within the range of values observed in previous work.

Chapter 5: Analysis 3

The results above show that the !Xung and Ju networks are less resilient to phonotactic disruption than the English network. Here, we investigate which phonotactic properties of these languages might be most important in maintaining their structure. Since bigram pseudolexicons do not show the important network properties which are preserved in trigram lexicons, we conclude that the phonotactic constraints responsible for the network structure are not strictly local. Thus, in Analysis 3, we construct pseudolexicons which preserve global constraints on the shapes of words in these languages. We compare these pseudolexicons within each language to gauge the importance each property has in creating network structure.

Methodology

Both !Xung and Ju allow syllables to begin with at most one consonant and do not allow codas. To highlight these phonotactics, we generate scrambled (as above) and CV (generate words using actual syllabic structure distribution) pseudolexicons for each of our corpora. Because !Xung and Ju words tend to begin with a click, for these languages we also generate semi-scrambled (scramble word in place but any present click stays at the initial position) and KCV (generate words using actual syllabic structure distribution with clicks being their own category) pseudolexicons. The pseudolexicons have the same word length distribution as the original lexicon. Examples of words from these pseudolexicons can be found in Table 3 within the appendix.

We compare the network properties of these pseudolexicons (averaged over 10 trials) within each language. We follow the procedure from the preliminary examination to create and analyze each LN.

Results

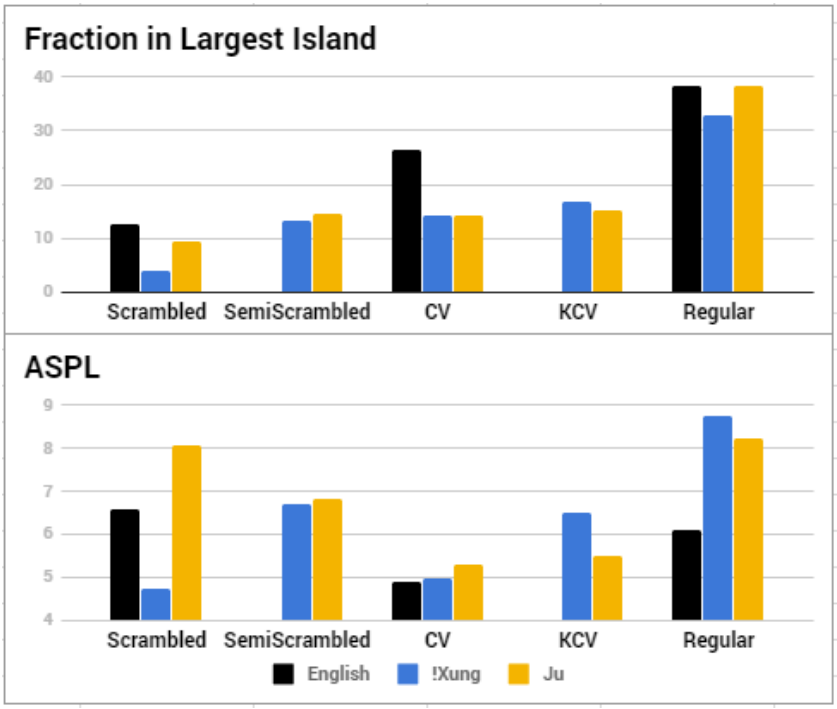

Figure 4: Network properties for English, !Xung, and Ju pseudolexicons which highlight phonotactic properties

We compare each languages' pseudolexicons within themselves to investigate the effects of each phonotactic property on network structure. Looking at Figure 4, we see there is a substantial increase in largest island size and ASPL from scrambled to semi-scrambled and CV. This highlights the importance of click position and syllabic structure. However, there is still a considerable disparity between KCV and natural language. This shows that combining both click position and syllabic structure does not necessarily add much structure, likely due to an overlap between categories. Also this indicated that there are

other important factors that determine structure which are not accounted for within syllabic structure and click location. In general pseudolexicons that preserve the characteristic positions of clicks, vowels and pulmonic consonants do a reasonable job of matching the network properties despite not including any local phone-to-phone dependencies. However, trigrams are closer yet, showing that other phonotactic patterns also play a role.

Chapter 6: Conclusion

Overall, we find that !Xung and Ju do not substantially differ in network properties when compared to European languages despite fundamental phonological disparities. This supports the argument of Vitevich (2008) that the LNPs indicate an underlying cognitive structure which is necessary for efficient word retrieval. Moreover, it suggests a selective pressure shaping the phonotactics of these languages (and others with large inventories) -- phonotactic rules may arise and change over time in ways that preserve the network properties within a cognitively useful range. For instance, the differences between randomly scrambled and syllabic pseudowords indicate that the restricted syllable inventories of !Xung and Ju may force words to cluster more tightly in the LN, compensating for the large number of contrastive phonemes. In other words, the underlying universal structure may be, not linguistic, but cognitive: the memory architecture responsible for word retrieval. This universal architecture may require certain patterns of connectivity within the lexicon, and these, in turn, may entail particular phonotactic patterns.

We are currently generating more pseudolexicons based on stricter phonotactics of !Xung and Ju, highlighting properties such as phoneme distribution within syllables. We hypothesize these further analyses work to present a more complete picture as to which phonotactics are critical in maintaining LNPs.

Looking forward, we plan to expand our current LN analysis to include languages with small phoneme inventories, such as certain Polynesian languages. Through this, we

hope to uncover how our hypothesis operates in languages starkly dissimilar from !Xung, Ju, and European languages.

Additionally, we plan to calculate functional load on !Xung and Ju. A LN assumes real world speakers can distinguish perfectly between minimal pairs however, with the large phoneme inventories of !Xung and Ju, these clicks may be confusable in real speech. As such, we are looking to employ functional load calculations, as done in Surendran and Niyogi (2006), on click words. We will use a trigram model at the word level to provide context for each word to measure the confusion between similar-sounding words by asking whether similar sounding words occur in the same context?

Bibliography

Biesele, M. , Boo, B. C., Gcao, H. K., G‡kao M. /K, Kaqece K N!.,, Miller, A., /Kunta, /A. F, Tsamkxao F. /U. /Ui, C. N! & (2006) Ju|'hoansi Dictionary, Revised version of Dickens, P. Ju|'hoan-English – English-Ju|'hoan Dictionary, Unpublished manuscript. The Kalahari People's Foundation and The Ju|'hoan Transcription Group.

Bradfield, J. (2014). Clicks, concurrency and Khoisan. Phonology,31(01),1-49. doi:10.1017/s0952675714000025

David, C. C., Miller, D., & Walker, K. (2004). The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text. In 4th International Conference on Language Resources and Evaluation (pp. 69-71). Lisbon, Portugal: LREC.

Dickens, P. (1994). English - Juhoan, Juhoan - English dictionary. Köln: Köppe.

Gruenenfelder, T. M., & Pisoni, D. B. (2009). The Lexical Restructuring Hypothesis and Graph Theoretic Analyses of Networks Based on Random Lexicons. Journal of Speech Language and Hearing Research, 52(3), 596.

Heafield, K. (2011, July). KenLM: Faster and smaller language model queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation (pp. 187-197). Association for Computational Linguistics.

Maddieson, I. (2013). Consonant Inventories.In: Dryer, Matthew S. & Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Maddieson, I. (2013). Vowel Quality Inventories. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Miller, A. L. (2016). Posterior lingual gestures and tongue shape in Mangetti Dune !Xung clicks. Journal of Phonetics, 55, 119-148.

Miller, A., Namaseb, L., Sands, S., Shah, S. Aromo, M., Augumes, C., Fransisko, R., Kaleyi, T., Prata, D., Riem, S., (2008). Mangetti Dune !Xung Dictionary. Unpublished Manuscript. The Ju|'hoan Transcription Group, The Kalahari People's Foundation, The University of Namibia and Northern Arizona University.

Miller-Ockhuizen, A. and Zec, D. (2003). Phonetics and Phonology of Contrastive Palatal affricates. Working Papers of the Cornell Phonetics Laboratory 2003, v.15, pp.130-193

Newman, M., & Girvan, M. (2003). Mixing Patterns and Community Structure in Networks. Statistical Mechanics of Complex Networks Lecture Notes in Physics, 66-87.

Sands, Bonny. (2010). Juu Subgroups Based on Phonological Patterns. In Brenzinger, Matthias and König, Christa (eds.), Khoisan Language and Linguistics: the Riezlern Symposium 2003, 85-114. Cologne: Rüdiger Köppe.

Shoemark, P., Goldwater, S., Kirby, J., & Sarkar, R. (2016). Towards robust cross-linguistic comparisons of phonological networks. Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology.

Stella, M., & Brede, M. (2015). Patterns in the English language: phonological networks, percolation and assembly models. Journal of Statistical Mechanics: Theory and Experiment, 2015(5).

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. Proceedings of Interspeech

Stolcke, A., Zheng, J., Wang, W., & Abrash, V. (2011). SRILM at sixteen: Update and outlook. Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (Vol. 5).

Surendran, D., and Niyogi, P. (2006). Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. In O. N. Thomsen, ed., Competing Models of Linguistic Change: Evolution and Beyond. Amsterdam and Philadelphia: John Benjamins.

Turnbull, R., & Peperkamp, S. (2016). What governs a language's lexicon? Determining the organizing principles of phonological neighbourhood networks. Studies in Computational Intelligence Complex Networks & Their Applications V, 83-94.

Vitevitch, M. S. (2008). What Can Graph Theory Tell Us About Word Learning and Lexical Retrieval? Journal of Speech Language and Hearing Research, 51(2), 408

Appendix

Table 1 : LNPs of !Xung and Ju. The labels in the parentheses note which language had the most similar results according to Shoemark et al. (2016).

Language	!Xung (974 Words)	Ju (3733 Words)
ASPL	8.756 (German)	8.211 (German)
Deg. Assortativity	0.541 (German)	0.709 (French)
Clustering Coefficient	0.511 (French/Polish)	0.422 (French/Polish)
Frac. Largest Island	32.752 (Dutch)	38.253 (English)

Table 2 : Degree statistics of !Xung and Ju

Language	!Xung (974 Words)	Ju (3733 Words)
Median Deg.	4.0	5.0
Mean Deg.	4.357	5.994
Min Deg.	1	1
Max Deg.	14	24
Deg. Std. Dev.	2.749	4.379

Table 3 : Examples of 3 “words” from each pseudolexicon

	English	!Xung	Ju
Uniform	ŋsθ, iɛhh, iʒgmʊ	ǜǂʰë̯, qʒò, ȅȅa̋aˀ	pònǂn!é, gǂ!ǂù'n, fSgǂgǂAeli'
Zipfian	oʊpə, ɔəɛŋə, ŋɔæɔs	ómǹ|n̊|ʰ, g‖ŋ!̀, g|hǫ̀	m'ù'y‖'n!g, ‖'om'uò, ò'm'‖'i'm'
Scrambled	ɛθlih, əwrdnəɪŋ, krljəɛaən	ə̀‖ŋ, eŋ̊!ʰe, a̋àʰ!	àkphù, aǂuh, gà's‖íí
Semi-Scrambled	(N/A)	‖ŋə̀, ŋ̊!ʰee, !àʰa̋	pùhàk, ǂauh, í‖ísgà'
CV	ɛnfəzamɛs, vɪlɔkitdr, nnaəln	nə̰̀!‘úŋ|, g‖áχúβ, ʃínòʒ	hà!écan, gm|ò!'h‖àq'à, ozmú
KCV	(N/A)	ŋ!dóm, ǂ’zús’, g‖lòs	!'hùúnlà, ǂèánn, n!aòkt
Unigram	wjork, sɛgit, njŋe	úŋ̀‖úlùb, téùχàˢà, íl‖!ìò	kéúháòno, à!'í‖nhàs, xàn!!gsat
Bigram	rnɔsm, ɔjə, tekjks	!ʔḿ, màˢíí, l’ùβa̤ⁿ	!'hármà, ǂ'haùbéá, n!h
Trigram	plæŋ, lɪŋ, hæv	ŋ̊|ʰùì, ŋ̊‖ʰù̀là, tà̋ʰ	‖'há, nǂuùnín, páróséré
Natural Lexicon	hɛlθi, wəndərɪŋ, kɛrəlajnə	‖ə̀ŋ, ŋ̊!ʰee, !a̋àʰ	phùkà, ǂauh, g‖à'ísí